The Education Policy Center
AT MICHIGAN STATE UNIVERSITY

WORKING PAPER #18

Can Value-Added Measures of Teacher Performance Be Trusted?

Cassandra M. Guarino, Ph.D.
Indiana University Bloomington

Mark D. Reckase, Ph.D.
Jeffrey M. Wooldridge, Ph.D.
Michigan State University

May 24, 2012
Note: This paper was originally posted March 2011, revised April 2011 and May 2012

Can Value-Added Measures of Teacher Performance Be Trusted?

Author Note

This work was supported by grant no. R305D10002 from the Institute of Education Sciences in the U.S. Department of Education. We thank Francis Smart for programming assistance, Brian Stacy and Eun Hye Ham for research assistance, and Steven Haider, Ken Frank, Anna Bargagliotti, Steven Dieterle, Brian Stacy, Francis Smart, Cory Koedel, Doug Harris, Jacob Vigdor, and Jonah Rockoff for helpful comments.

Abstract

We investigate whether commonly used value-added estimation strategies can produce accurate estimates of teacher effects. We estimate teacher effects in simulated student achievement data sets that mimic plausible types of student grouping and teacher assignment scenarios. No one method accurately captures true teacher effects in all scenarios, and the potential for misclassifying teachers as high- or low-performing can be substantial. Misspecifying dynamic relationships can exacerbate estimation problems. However, some estimators are more robust across scenarios and better suited to estimating teacher effects than others.

Can Value-Added Measures of Teacher Performance Be Trusted?

By Cassandra M. Guarino, Mark D. Reckase, Jeffrey M. Wooldridge

May 17, 2012

Cassandra M. Guarino
Associate Professor of Educational Leadership and Policy Studies
Indiana University Bloomington
Email: guarino@indiana.edu

Mark D. Reckase
University Distinguished Professor of Measurement and Quantitative Methods
Michigan State University
Email: reckase@msu.edu

Jeffrey M. Wooldridge
University Distinguished Professor of Economics
Michigan State University
Email: wooldri1@msu.edu

Acknowledgment: This work was supported by grant no. R305D10002 from the Institute of Education Sciences in the U.S. Department of Education. We thank Francis Smart for programming assistance, Brian Stacy and Eun Hye Ham for research assistance, and Steven Haider, Ken Frank, Anna Bargagliotti, Steven Dieterle, Brian Stacy, Francis Smart, Cory Koedel, Doug Harris, Jacob Vigdor, and Jonah Rockoff for helpful comments.

Abstract: We investigate whether commonly used value-added estimation strategies can produce accurate estimates of teacher effects. We estimate teacher effects in simulated student achievement data sets that mimic plausible types of student grouping and teacher assignment scenarios. No one method accurately captures true teacher effects in all scenarios, and the potential for misclassifying teachers as high- or low-performing can be substantial. Misspecifying dynamic relationships can exacerbate estimation problems. However, some estimators are more robust across scenarios and better suited to estimating teacher effects than others.

1. Introduction

Accurate indicators of educational effectiveness are needed to advance national policy goals of raising student achievement and closing socioeconomically based achievement gaps. If constructed and used appropriately, such indicators for both program evaluation and the evaluation of teacher and school performance could have a transformative effect on the nature and outcomes of teaching and learning. Measures of teacher quality based on value-added models of student achievement (VAMs) are gaining increasing acceptance among policymakers as a possible improvement over conventional indicators, such as classroom observations or measures of educational attainment or experience. They are already in use to varying degrees in school districts[1] and widely reported in the research literature.

Intuitively, VAMs are appealing; they track growth in learning from one year to the next for individual students and parse that growth into pieces believed to represent the separate contributions made by teachers and schools as well as individual-specific factors. Moreover, given that standardized testing is now ubiquitous in U.S. school systems, VAMs can be inexpensive to implement relative to other forms of teacher evaluation such as classroom observation, and their use has been encouraged by Race to the Top (U.S. Department of Education, 2009). As a teacher evaluation tool, VAM-based measures are sometimes viewed as less subjective than judgments based on observations by principals or portfolios of accomplishments. Given the increasing visibility of VAM-based estimates of teacher and school quality, and the possible inclusion of teacher performance incentives in the upcoming reauthorization of NCLB, it is imperative that such measures be well constructed and understood.

Controversy exists, however, as to the best way to construct VAMs and to their optimal application. Numerous methods have been developed (e.g., Sanders & Horn, 1994; Ballou,

[1] In some districts, the popular press has computed and published teacher value-added scores. For example, in September 2010, the Los Angeles Times, after analyzing data obtained from Los Angeles Unified School District officials under California's Public Records Act, created a website in which any member of the public can look up VAM-based ratings for individual public school teachers in grades 3 through 5. See: http://www.latimes.com/news/local/teachers-investigation/ (downloaded 10/12/10). In New York City, after a protracted court battle, the district has been required to make teacher evaluation measures available to the public. See http://www.nytimes.com/2012/02/25/education/teacher-quality-widely-diffused-nyc-ratings-indicate.html (downloaded 5/4/12).

Sanders, & Wright (2004); McCaffrey et al., 2004; Kane & Staiger, 2008; Raudenbush, 2009), and studies that compare estimates derived from different models have found substantial variability across methods (McCaffrey et al., 2004). Concerns remain that our understanding of these models is as yet limited and that incentives built around them may do more harm than good, with teachers' unions, in particular, reluctant to allow their constituents to be judged on the basis of measures that are potentially biased or imprecise.

A central issue involved in establishing the validity of measures and inferences based on VAMs is whether VAMs effectively isolate the "true" contribution of teachers and schools to achievement growth or instead confound these effects with the effects of other factors that may or may not be within the control of teachers and schools. Given that neither students nor teachers are randomly assigned to schools and that students are not randomly assigned to teachers within schools, disentangling the causal effects of schooling from other factors influencing achievement is far from straightforward. The few studies that have attempted to validate VAMs have drawn different conclusions (e.g., Kane & Staiger, 2008; Rothstein, 2010)[2], and questions about the validity of VAMs linger.

In this paper, we investigate the ability of various estimation strategies to produce accurate estimates of teacher effects. Our main research question is the following: How well do commonly used estimators perform in estimating teacher effects under a variety of known conditions, including those in which particular underlying assumptions are violated?

We focus our study on estimators that are commonly used in research and policy applications involving teacher effects. We first outline the assumptions that must be met for each estimator to have desirable statistical properties in the context of a conventional theoretical framework. We then apply the estimators to the task of recovering teacher effects in simulated student achievement data generated under different types of student grouping and teacher assignment scenarios. We then compare the estimated teacher effects to the true teacher effects embedded in the data.

The paper is organized as follows. In Section 2, we outline a theoretical framework for value-added models based on a well-known structural cumulative effects model. Section 3

[2] Kane and Staiger (2008) compare experimental VAM estimates for a subset of Los Angeles teachers with earlier non-experimental estimates for those same teachers and find that they are similar, suggesting that they are valid. Rothstein (2008), on the other hand, devises falsification tests that challenge the validity of VAM-based measures of teacher performance in North Carolina.

discusses each estimator in turn and its underlying assumptions. An important component of our study is that we apply all estimators to all of our simulation conditions—even those in which they are not necessarily expected to perform well in order to determine whether some methods are more robust than others across different scenarios. We describe different mechanisms for grouping students and assigning teachers to classrooms in Section 4. Section 5 describes the simulation procedures and estimation strategies we employ. The simulations results in Section 6 investigate the ability of the various value-added estimators of teacher performance to uncover true effects under our different data generating scenarios. By systematically comparing VAM-based estimates resulting from different estimators to the true effects embedded in the various data generating processes, we are able to identify estimation strategies most likely to recover true effects under specific conditions.

Our investigations yield several important findings. No one estimator performs well under all plausible circumstances, but some are more robust than others. Surprisingly, certain estimation approaches known to be inconsistent in the structural modeling framework fare better than expected. Our simulations highlight the pitfalls of misspecifying the dynamic relationship between current and prior achievement. In addition, we find that substantial proportions of teachers can be misclassified as "below average" or "above average" as well as in the bottom and top quintiles of the teacher quality distribution, even in the best-case scenarios.

An important caveat to apply to our findings is that they result from data generation processes that incorporate many of the assumptions underlying a relatively simple conceptual model. Thus, we subject the estimators to idealized conditions. Undoubtedly real-life educational conditions are more complex, and the estimators will likely perform less well when applied to real data. Detecting the flaws in various estimators under idealized conditions, however, is the best way to discover fundamental differences among them. Thus, the simplifications built into our research design are the strength of the design.

2. A Common Approach to Value-Added Modeling

The derivation of particular VAMs typically rests on the specification of a structural "education production function," in which achievement at any grade is modeled as a function of child, family, and schooling inputs. In its most general formulation, learning is a process that is considered to be both dynamic and cumulative – that is, past experiences and past learning contribute to present learning. Thus the model—often referred to as the generalized *cumulative*

effects model (CEM)—includes all relevant past child, family, and school inputs (Hanushek, 1979, 1986; Boardman & Murnane, 1979; Todd & Wolpin, 2003; Harris, Sass, & Semykina, 2011). This model can be expressed as:

$$A_{it} = f_t(E_{it},\ldots, E_{i0},X_{it},\ldots,X_{i0},c_i, u_{it}) \quad (1)$$

where A_{it} is the achievement of child i in grade t, E_{it} represents school-related inputs, X_{it} represents a set of relevant time-varying child and family inputs, c_i captures the unobserved time-invariant student effect (representing, for example, motivation, some notion of sustained ability, or some persistent behavioral or physical issue that affects achievement), and the u_{it} represent the idiosyncratic shocks that may occur in any given period. In this very general formulation, the functional form is unspecified and can vary over time.

Moving to an empirical model poses large challenges due to the lack of information regarding most past and even many current inputs to the process and the manner in which they are related to one another—that is, functional form, interactions, lagged responses, feedback, and so on. Inferring the causal effects of teachers and schools is therefore difficult. If children were randomly assigned to teachers and schools, many omitted variable issues would be considerably mitigated. However, random assignment does not typically characterize school systems, and, indeed, is not necessarily desirable. Random assignment of children to schools deprives parents of the ability to find schools that they believe to be best suited for their children through both residential sorting and school choice. Random assignment to teachers within schools deprives principals of one of their most important functions: to maximize overall achievement by matching the individualized skills of teachers to those students most likely to benefit from them. Thus random assignment—while helpful from an evaluation standpoint—could result in suboptimal learning conditions if particular teacher and school characteristics interact in a beneficial way with student characteristics in the learning process.

Clearly, however, knowledge of the effectiveness of particular schools, teachers, or programs in promoting learning is essential if we are to foster successful instructional approaches and curtail the use of ineffective ones. Causal measures of performance at the school, teacher, and program level are needed to identify instructional strategies that contribute to high performance. In the context of nonrandom assignment and omitted variables, statistical methods are the only tools available with which to infer effects, but they rely on strong assumptions. In

the next sections, we describe the assumptions used to derive models that are empirically feasible to estimate.

2.1. The General Linear Formulation

A distributed lag version of the cumulative effects model that assumes linearity is the typical and potentially tractable starting point for structural modeling. Equation (1) becomes

$$A_{it} = \alpha_t + E_{it}\beta_0 + E_{i,t-1}\beta_1 + \ldots + E_{i0}\beta_t + X_{it}\gamma_0 + X_{i,t-1}\gamma_1 + \ldots + X_{i0}\gamma_t + \eta_t c_i + u_{it} \qquad (2)$$

where we take E_{it} to be a row vector of observed education inputs at time t – including teacher or school characteristics, or, say, teacher indicators – and X_{it} to be a vector of observed time-varying individual and family characteristics such as health status, household income, and so on. The term α_t allows for a separate intercept in each time period, which would be appropriate if, for example, the score scales are set to be different for different grade levels by the testing program. The period $t = 0$ corresponds to the initial year in school (which is generally kindergarten or could be pre-kindergarten in states where this is a common public school option). This formulation has several assumptions embedded in it: linearity, a functional form that is constant over time (except for the intercept and possibly the coefficient on c_i), and an additive, idiosyncratic shock, u_{it}, that accounts for all unobserved time-varying current and past factors.

Note that the formulation in (2) does not explicitly recognize the possible presence of interactions among teachers, between teachers and students, or among students and is therefore a limited conceptualization of the educational learning process. It is possible to build in these complexities, although it is rarely done in practice, except for the occasional inclusion of peer characteristics.

Exogeneity assumptions on the inputs are needed to estimate the parameters in the linear CEM. A common starting point—termed *sequential exogeneity* by Chamberlain (1992)—assumes that the expected value of the time-varying unobservables, u_{it}, conditional on all relevant time-varying current and past inputs and the unobserved child effect, is zero:

$$E(u_{it} | E_{it}, E_{i,t-1}, \ldots, E_{i0}, X_{it}, X_{i,t-1}, \ldots, X_{i0}, c_i) = 0 \quad . \qquad (3)$$

In practical terms, (3) requires that the time-varying unobservables that affect achievement are uncorrelated with observed school and family inputs—both current and past. Initially, this may seem reasonable given that most school input decisions, such as teacher and class size assignments, are made at the end of the previous school year, and cannot be based on changes in a student's situation over the course of the school year. However, u_{it} can contain factors such as

unobserved parental effort that respond to the assignment of school inputs. For example, a parent may provide more help for a student who is assigned to a poor teacher or a large class.

As stated, (3) is an assumption about correlation between inputs and the time-varying unobservables, u_{it}, and it does not address the relationship between student heterogeneity, c_i, and the observed inputs. Many estimation approaches either ignore the presence of c_i or assume it is uncorrelated with observed inputs – in other words, they assume what we would call *heterogeneity exogeneity*. If, however, c_i is correlated with observed inputs, then standard pooled regression and generalized least squares approaches are generally inconsistent regardless of what we assume about the relationship between u_{it} and the inputs. Several approaches can be used to deal with unobserved heterogeneity in equation (2)—most commonly, fixed effects and first-differencing methods—each with a set of assumptions and drawbacks. If we are not wedded to a structural model as in equation (2), past test scores can be included in regression equations as proxies for the unobserved heterogeneity. In fact, this is an important motivation for the dynamic regression method described in Section 3.

Beyond the issue of unobserved heterogeneity, there are other obstacles to estimating equation (2). The linear CEM in this form is rarely estimated due to data limitations. If, for example, we have testing data on 3rd through 6th grade for each child and want to allow for the possibility that all previous teachers affect current outcomes (in this case, the E_{it} vector may be composed of teacher dummy variables), we need to have data linking students to their teachers in 2nd and 1st grades, as well as kindergarten. In addition to the onerous data requirements, high correlations among inputs across time periods can limit the ability of any of these estimators to isolate specific contemporaneous or past effects and make estimation of the linear CEM unattractive.

2.2. Geometric Distributed Lag Restrictions on the Linear CEM

To solve the data limitations issue and conserve on parameters in the general linear CEM, researchers typically impose restrictions on the distributed lag coefficients. A simple and commonly applied restriction is a *geometric* distributed lag (GDL), which imposes geometric decay on the parameters in (2) for some $0 \leq \lambda \leq 1$:

$$\beta_s = \lambda^s \beta_0, \ \gamma_s = \lambda^s \gamma_0, \quad s = 1,\ldots,T \tag{4}$$

This means that the effects of all past time-varying inputs (schooling-related as well as child- and family-related) decay at the same rate over time and their influence on current achievement

decreases in the specified manner as their distance from the present increases. With these restrictions, after subtracting $\lambda A_{i,\,t-1}$ from both sides of (2) and performing substitutions and simple algebra, we obtain a much simpler estimating equation:

$$A_{it} = \tau_t + \lambda A_{i,t-1} + E_{it}\beta_0 + X_{it}\gamma_0 + \pi_t c_i + e_{it} \qquad (5)$$

where

$$e_{it} = u_{it} - \lambda u_{i,t-1} \qquad . \qquad (6)$$

Equation (5) has several useful features. First, the right hand side includes a single lag of achievement and only contemporaneous inputs. This is a much more parsimonious estimating equation than the general model (2) because past inputs do not appear. Consequently, data requirements are less onerous than those for the linear CEM, and parameter estimation of (5) is less likely to suffer from the multicollinearity that can occur among contemporaneous variables and their lags.

It is important to see that the decay structure in the GDL equation means that any distributed lag effects are determined entirely by λ and β_0. In other words, once we know the effect of contemporaneous inputs (β_0) and the decay parameter (λ), the effects of lagged inputs are determined. Undoubtedly this is a highly restrictive assumption, but (5) is fairly common in the education literature. It is important to note, however, that the rate at which knowledge decays may differ for different students or for different subpopulations of students (Entwistle & Alexander, 1992; Downey, Hippel & Broh, 2004). Although allowing rates of decay to vary by individuals or groups is possible in (5), this is rarely, if ever, done in the literature on teacher effects.

In deriving estimators based on equation (5), we must consider the exogeneity of inputs in this equation, including possible correlation with c_i as well as correlation with the time-varying unobservables e_{it}. As shown in equation (6), e_{it} depends on the current and lagged error from equation (2). If we maintain the sequential exogeneity assumption (3) in the structural CEM, u_{it} is uncorrelated with E_{it}. In that case, simple algebra gives

$$Cov(E_{it}, e_{it}) = -\lambda\, Cov(E_{it}, u_{i,t-1}). \qquad (7)$$

Equation (7) shows explicitly that in order to treat E_{it} and X_{it} as exogenous in (5) – that is, uncorrelated with the time-varying unobservables e_{it} – we need to impose an assumption stronger than the sequential exogeneity in the structural equation (2) (unless $\lambda = 0$, which seems unlikely). In this case, the weakest exogeneity condition is that E_{it} is uncorrelated with $u_{it} - \lambda u_{i,\,t-1}$, which

could be true even if we do not assume E_{it} is uncorrelated separately with $u_{i,\,t-1}$ and u_{it}. However, for certain estimation strategies discussed below, the imposition of a stronger exogeneity assumption on the CEM, namely *strict exogeneity*, is needed and is clearly sufficient for Cov(E_{it}, e_{it}) = 0. A straightforward way to state the strict exogeneity assumption is

$$E(u_{it} | E_{iT}, E_{i,T-1},\ldots, E_{i0,} X_{iT}, X_{i,T-1},\ldots, X_{i0,} c_i) = 0. \qquad (8)$$

The difference between assumptions (8) and (3) is that (8) includes the entire set of observed inputs, including *future* inputs (this is why the *t* in (3) is replaced with *T* in (8)). Assumption (8) implies that the error term e_{it} in (5) is uncorrelated with inputs at time *t* and all other time periods.

In addition to possible correlation between the covariates and e_{it}, we must recognize that it is virtually impossible for c_i to be uncorrelated with $A_{i,\,t-1.}$ Moreover, we often expect c_i to be correlated with the inputs.

A simplistic approach to dealing with issues stemming from the presence of the lagged dependent variable is to assume that it does not matter – that is, assume that $\lambda = 0$, which implies complete decay. In this case, (5) reduces to what is often referred to as a "level-score" equation. As a special case of the CEM, the level-score approach is unattractive because $\lambda = 0$ is unrealistic. But level-score regressions have been used with experimental data – that is, when the inputs are randomly assigned – because then the structural CEM approach is not necessary for identifying teacher effects (see, for example, Dee, 2004). For estimating teacher value added, random assignment means that one can compare mean achievement scores across teachers, and that is exactly what level-score regressions do in that setting.

Another simple but very widely used formulation sets $\lambda = 1$ (no decay), which leads to subtracting $A_{i,\,t-1}$ from both sides of (5), thereby achieving a so-called "gain score" formulation:

$$\Delta A_{it} = \tau_t + E_{it}\beta_0 + X_{it}\gamma_0 + \pi_t c_i + e_{it}. \qquad (9)$$

We now turn to describing different estimators used to estimate VAMs along with their statistical properties.

3. Commonly Used Estimators and their Underlying Assumptions

This section describes six commonly used estimation methods and the assumptions underlying their use. One caveat regarding our discussion of assumptions is that we appeal to large-sample properties because several of the estimators have no tractable finite-sample properties (such as unbiasedness) under any reasonable assumptions. Appealing to asymptotic

analysis is hardly ideal, especially for applications where the inputs are teacher assignments. In this scenario, the large-sample approximation improves as the number of students per teacher increases. But in many data sets, the number of students per teacher is somewhat small – fewer than 100 – making large-sample discussions tenuous. Nevertheless, asymptotic theory is the unifying theme behind the estimators that are applied in VAM contexts and provides a framework within which to identify underlying assumptions.

3.1. Dynamic Ordinary Least Squares

If we write equation (5) with a composite error v_{it}, as

$$A_{it} = \tau_t + \lambda A_{i,t-1} + E_{it}\beta_0 + X_{it}\gamma_0 + v_{it}, \qquad (10)$$

and ignore the properties of v_{it} – that it depends on $\pi_t c_i$ and (the possibly serially correlated) e_{it} – then we might take a seemingly naïve approach and simply estimate a dynamic regression. In other words, we might estimate λ, β_0, and γ_0 using a pooled OLS regression. We will refer to this estimator as "dynamic ordinary least squares" (DOLS).

Consistency of the DOLS estimator for β_0, and γ_0 (and λ)—which, recall, are parameters in the structural model—hinges on strict exogeneity of the inputs (with respect to $\{u_{it}\}$) and no serial correlation in $\{e_{it}\}$. Since $e_{it} = u_{it} - \lambda u_{i,t-1}$, to claim that the $\{e_{it}\}$ are serially uncorrelated, we must place restrictions on the original errors $\{u_{it}\}$. First, we must assume they follow an *AR(1)* process, namely $u_{it} = \rho u_{i,t-1} + r_{it}$ where $\{ r_{it} \}$ is serially uncorrelated, and, second, we must assume that $\rho = \lambda$., which is often called the "common factor" (CF) restriction. The CF restriction amounts to assuming that past shocks to learning decay at the same rate as learning from family- and school-related sources. This is by no means an intuitive assumption. In any case, under the CF restriction the transformed errors $e_{it} = u_{it} - \lambda u_{i,t-1}$ in (5) are the same as the serially uncorrelated r_{it}.

In addition, the presence of $\pi_t c_i$ generally causes inconsistency because c_i is correlated with $A_{i,t-1}$ and possibly the inputs E_{it}, too, which will be the case if students are assigned to educational inputs based on time-constant unobservables. Controlling for background variables can mitigate the problem, but proxies for c_i may be hard to come by, and those easily available (for example, gender or race) are likely insufficient to proxy motivation or persistent correlates of ability.

Even if, technically speaking, DOLS is inconsistent, it could nevertheless provide relatively accurate estimates of β_0 under certain circumstances. For example, if the $\pi_t c_i$ are

sufficiently "small," ignoring this component of the composite error term v_{it} might not be costly. Even with substantial heterogeneity, the lagged test score may serve as a good proxy for c_i, resulting in good estimators of β_0 even though λ may be poorly estimated. An attractive feature of DOLS is that controlling for $A_{i,t-1}$ explicitly allows for the kinds of dynamic assignment of students to inputs based on prior test scores.

3.2. The Arellano and Bond Approach

Rather than ignore c_i, a combination of first differencing and instrumental variables can be used to account for unobserved heterogeneity, again assuming that π_t is a constant. We can eliminate c_i by first differencing (5) to obtain:

$$\Delta A_{it} = \chi_t + \lambda \Delta A_{i,t-1} + \Delta E_{it}\beta_0 + \Delta X_{it}\gamma_0 + \Delta e_{it}. \quad (11)$$

Generally, this differenced equation cannot be consistently estimated by OLS because $\Delta A_{i,t-1}$ is correlated with Δe_{it}. Nevertheless, under strict exogeneity of inputs $\{E_{it}\}$ and $\{X_{it}\}$, Δe_{it} is uncorrelated with inputs in any time period, and so it is possible to use lagged values of E_{it} and X_{it} as instrumental variables for $\Delta A_{i,t-1}$. (ΔE_{it} and ΔX_{it} act as their own instruments under strict exogeneity.) If we use more than one lag – as is often required to make the instruments sufficiently correlated with the changes – this IV approach increases the data requirements because we lose an additional year of data for each lag we include among the instruments. For example, if we use the lagged changes, $\Delta E_{i,t-1}$ and $\Delta X_{i,t-1}$, as IVs, we lose one year of data because these depend on $E_{i,t-2}$ or $X_{i,t-2}$, respectively. Thus, this estimator is rarely applied in practice. Instead, the estimator proposed by Arellano and Bond (1991) (AB), which chooses instruments for the lagged gain score from available achievement lags, is more often used (e.g., Koedel & Betts, 2011).

The AB approach is limited by its requirement that there be no serial correlation in the $\{e_{it}\}$, thus imposing the common factor restriction described above. Formally stated, an assumption that implies no serial correlation in the errors and strictly exogenous inputs is:

$$E(e_{it} | A_{i,t-1}, A_{i,t-2}, \ldots, A_{i0}, E_{iT}, E_{i,T-1}, \ldots, E_{i0}, X_{iT}, X_{i,T-1}, \ldots, X_{i0}, c_i) = 0, \quad (12)$$

which maintains that e_{it} is unpredictable given past achievement and the entire history of inputs. The usefulness of assumption (12) is that it implies that $\{A_{i,t-2}, \ldots, A_{i0}\}$ are uncorrelated with e_{it}, and so these are instrumental variable candidates for $\Delta A_{i,t-1}$ in (11). Typically, $\{A_{i,t-2}, \ldots, A_{i0}\}$ is sufficiently correlated with $\Delta A_{i,t-1}$, as long as λ is not "close" to one. With achievement scores

for four grades, and teacher assignments for the last three, equation (11) can be estimated using two years of gain scores.

Generally, care is needed when instrumenting for $\Delta A_{i,t-1}$ when λ is "close" to one. In fact, if there were no inputs and $\lambda = 1$, the AB approach would not identify λ. Simulation evidence in Blundell and Bond (1998) and elsewhere verifies that the AB moment conditions produce noisy estimators of λ when λ is near one. We should remember, though, that our main purpose here is in estimating school input effects (in our case, teacher effects), β_0, rather than λ. For that purpose, the weak instrument problem when λ is near unity may not cause the AB approach to suffer too severely.

If we wish to allow for the possibility of dynamic assignment and *not* assume strict exogeneity of the inputs in (2), then ΔE_{it} requires instruments as well, and this is a tall order. In (11), Δe_{it} depends on $\{u_{it}, u_{i,t-1}, u_{i,t-2}\}$ and so, if we hope to relax strict exogeneity of the inputs in (2), we must choose our IVs from $\{A_{i,t-2}, \ldots, A_{i0}, E_{i,t-2}, \ldots, E_{i0}, X_{i,t-2}, \ldots, X_{i0}\}$. This approach imposes substantial data requirements.

3.3. Pooled OLS on the Gain Score

Estimation based on equation (9), where the gain score, ΔA_{it}, is used as the dependent variable and contemporaneous inputs are the explanatory variables is advantageous if the assumption that $\lambda = 1$ holds. If we can ignore the presence of c_i or successfully introduce proxies for it, pooled OLS (POLS) is a natural estimation method and is used in various applications (e.g., Ballou, Sanders & Wright, 2004).

A more subtle point is that when we view (9) as an estimating equation derived from the structural model (2), consistency of POLS relies on the same kind of strict exogeneity assumption we discussed in connection with (7): assignment of inputs at time t, E_{it}, cannot be correlated with the time-varying factors affecting achievement at time $t - 1$, $u_{i,t-1}$. If the inputs are strictly exogenous in the CEM then E_{it} is uncorrelated with e_{it}, and POLS is consistent provided the inputs are uncorrelated also with the unobserved heterogeneity. Inference for pooled OLS that allows arbitrary serial correlation and heteroskedasticity in the composite error $\pi_t c_i + e_{it}$ is straightforward.

When there are only teacher dummies E_{it} in equation (9) the POLS estimator is the same as computing the average gain score for each teacher.

3.4. The Average Residual Approach

When covariates other than teacher dummies appear in (9) – including past test scores – an approach to computing teacher performance measures that is widely implemented in the research literature (e.g., Chetty et al. 2011, McCaffrey et al. 2010, West & Chingos 2010, Kane & Staiger 2008) and for evaluation purposes in districts in various states (see the Wisconsin VARC estimator, Value-Added Research Center 2010) is based on the use of student-level residuals averaged at the teacher level. In the first step the gain score is regressed on the covariates (other than the teacher dummies) to obtain residuals. In the second step those residuals are averaged within teacher to obtain the teacher VAMs. The second step is equivalent to a pooled OLS regression of the residuals on teacher dummies. We use the acronym AR to describe this approach. The popularity of AR seems to mainly hinge on the computational simplicity of obtaining average residuals by teacher.[3] An important drawback to the AR approach, one which seems to have largely gone unnoticed, is that any correlation between teacher effects and included regressors, such as lagged test scores, is not purged in the first regression, thus generally leading to bias and inconsistency when such correlation exists. By contrast, DOLS properly accounts for correlation between E_{it} and $A_{i,t-1}$ in equation (10).

3.5. Random Effects on the Gain Score

A drawback to POLS – again assuming for the moment that $\lambda = 1$, the inputs are strictly exogenous, and the inputs are uncorrelated with student heterogeneity – is that it is generally inefficient in estimating β_0, because it ignores the serial correlation and heteroskedasticity in the composite error, $\pi_t c_i + e_{it}$. If we assume π_t is constant and that $\{e_{it}\}$ is serially uncorrelated and homoskedastic in equation (9), then random effects (RE) estimation can improve efficiency over POLS. Like POLS, RE assumes the heterogeneity is uncorrelated with inputs, but RE is guaranteed to be the efficient generalized least squares estimator when $\{e_{it}\}$ satisfies ideal assumptions.[4]

[3] The AR approach allows one to avoid running regressions with large sets of teacher dummies. In a cross-sectional setting, it also permits the inclusion of variables that do not change within classroom, such as classroom averages.

[4] When POLS and RE are both consistent, it should be noted that RE can still improve upon POLS in terms of efficiency even if $\{e_{it}\}$ is serially correlated or contains heteroskedasticity. Efficiency gains using RE in such settings are not guaranteed, but it is often more efficient that POLS because it accounts for serial correlation to some extent, even if not perfectly. This is the motivation behind the generalized estimating equations literature (see, for example, Zeger, Liang, & Albert 1988 or Wooldridge 2010, Chapter 1). Also, π_t not being constant does not cause inconsistency of RE (or POLS), although RE would not be the efficient GLS estimator with time-varying π_t. One

3.6. Fixed Effects on the Gain Score

If, instead of ignoring or proxying for c_i, we allow for unrestricted correlation between c_i and the inputs E_{it} and X_{it}, we can eliminate c_i in the gain score equation via fixed effects (FE) (at least when π_t is constant). The FE estimator also requires a form of strict exogeneity of E_{it} and X_{it} because FE employs a time-demeaning transformation that requires that the e_{it} are uncorrelated with the time-demeaned inputs.[5] As with the other methods, the strict exogeneity assumption stated in (8) is sufficient. When inputs related to classroom assignments are thought to be based largely on time-constant factors, FE is attractive, whereas POLS and RE will suffer from systematic bias. If inputs are uncorrelated with the shocks and heterogeneity, however, FE is typically less efficient than RE, and can be less efficient than POLS, too. Although FE is rarely used in practice to estimate teacher performance measures, we include it here for didactic purposes and note that it is sometimes used in value-added models designed to assess program effects at the teacher or school level.

3.7. Empirical Bayes and Related Estimators

A popular estimation approach to teacher VAMS is the so-called "empirical Bayes" (EB) method, application of which results in so-called "shrinkage estimators." The EB estimators are essentially the same as the VAMs obtained from the mixed model that is at the foundation of the Tennessee Value Added Assessment System (TVAAS) estimator originally developed by Sanders (for example, Ballou, Sanders & Wright 2004). Briefly, the teacher effects are modeled as random outcomes and then the best linear unbiased predictors are obtained as functions of the unknown variance parameters; estimates of the variance parameters are inserted to form the final shrinkage estimates. When applied to panel data, the EB approach loses its theoretical appeal in the presence of lagged test scores because the explanatory variables are no longer strictly exogenous. As is well known (for example, Morris, 1983), the EB VAM estimates when only teacher effects are included are simply the pooled OLS estimates shrunk toward the overall mean teacher effect, using a shrinkage factor that varies by class size. If class size is constant across teachers and time periods, as in the simulations we conduct, the EB estimator is identical (up to a scale factor) to various OLS estimators, such as the POLS and DOLS that we consider.

---

could instead use an unrestricted GLS analysis that would allow any kind of variance-covariance structure for $\pi_t c_i + e_{it}$. We do not explore that possibility in this paper, however, as it is rare in applications.

[5] For the same reason, a lagged dependent variable cannot be included on the right-hand side.

Consequently, the rankings of the teacher effects will be unchanged, and so we do not discuss the EB estimator separately in this paper.

3.8. Summary of Estimation Approaches

In summary, estimation of the parameters of the structural cumulative effects model, even after we impose the geometric distributed lag restriction to arrive at equation (5), requires numerous additional assumptions. OLS estimation of the dynamic equation – what we have called DOLS – requires strict exogeneity of inputs E_{it} and X_{it} and effectively imposes the common factor restriction on an AR(1) model for $\{u_{it}\}$. In addition, the method is not generally consistent if c_i is in the equation. In contrast to DOLS, the AB approach explicitly recognizes the presence of c_i but generally requires $\lambda < 1$, strict exogeneity of the inputs E_{it} and X_{it}, and no serial correlation in the $\{e_{it}\}$. POLS and RE on the gain score equation require strict exogeneity of inputs E_{it} and X_{it} and no correlation with c_i. FE allows for correlation between c_i and inputs E_{it} and X_{it} but maintains strict exogeneity. For either RE or FE to be an appropriate estimation method, however, λ must equal 1 (or a known value). The AR approach requires teacher effects to be uncorrelated with other covariates, whether those are other inputs, student background variables, or lagged test scores.

Violations of some assumptions may cause more severe problems in estimating teacher effects than violations of others, and it is thus an empirical question as to which estimator will perform best across various data generating mechanisms. For example, the AB approach is not guaranteed to dominate DOLS in every situation—if the inputs are not strictly exogenous, both estimators are technically inconsistent. Nor is AB necessarily better than approaches that impose $\lambda = 1$. At first glance it appears that (11) is more general because it does not impose $\lambda = 1$. But for AB to be consistent, serial correlation in the structural shocks $\{u_{it}\}$ must be of the AR(1) form and the CF restriction must hold. An important implication is that estimating λ when it is unity can be costly when using the first-differenced equation (11). In particular, if $\lambda = 1$ and the inputs are strictly exogenous, FE estimation of (9) consistently estimates the teacher effects without the CF restriction whereas AB estimation of (11) is generally inconsistent for the parameters in the CEM if the CF restriction fails. Because of this, we must be careful not to claim superiority of the AB approach over methods that do not require the CF restriction. The interesting question is which of the methods we have discussed does a better job recovering the coefficients on the inputs under different conditions, and that is what this study aims to answer.

4. Situating Theory in Context

Until now, we have discussed assumptions underlying value-added models and estimation strategies in relatively abstract terms. We now describe the types of educational scenarios that form the basis of our simulation design and how they might be expected to violate exogeneity assumptions.

We consider the process of matching students to teachers to be composed of two separate decisions—the grouping of students into classrooms and the assignment of classrooms to teachers. Grouping students in classrooms on the basis of their perceived ability, often called "tracking," is not uncommon and can take a number of forms. Students may be grouped together on the basis of either their prior test score, $A_{i,t-1}$,[6] their baseline level of achievement or ability upon entering school, A_{i0}, or their potential for learning gains, c_i. We will refer to the first type of ability grouping, a relatively common practice in educational settings as "dynamic tracking," following terminology used by Rothstein (2010). The second and third types of grouping, both forms of "static tracking," are less common. They might occur when, for example, schools either formally or informally assess the level of learning or the growth potential of children upon entering school, group the children accordingly, then keep more or less the same groups of children together for several grades.

Ability grouping in one form or another is likely to occur on a reasonable scale within educational systems. In the empirical literature, the phenomenon of ability grouping has been investigated primarily through techniques such as those developed by Aaronson, Barrow, and Sander (2007), which compare the average within classroom standard deviation in prior test scores with that produced by artificially regrouping students into classrooms—either randomly or perfectly sorted. Most such studies find that actual average standard deviations are closer to the random scenario than the perfectly sorting scenario. Dieterle et al. (unpublished), however, use a more fine-grained approach to the analysis of achievement data and find that substantial numbers of schools engage in dynamic tracking, a fact that can easily be obscured by the aggregated statistics.

Tracking does not, in and of itself, induce correlation between unobserved factors affecting student performance and teacher effects but serves as a precondition for the possibility

[6] Here for simplicity we refer to just one prior test score. However, principals might average over a series of prior test scores.

of correlation. We distinguish the practice of tracking— grouping of students together on the basis of some performance or ability criterion—from the practice of assigning these groups of students to teachers in nonrandom ways. In this study, we use the term "grouping" for the practice of placing students in classrooms and the term "assignment" for the action of assigning students to teachers.

In our study, assignment of classrooms to teachers take three primary forms: random assignment, assignment in which there is a positive correlation between teacher effects and student performance (that is, when better students are assigned to better teachers), and assignment in which there is a negative correlation between teacher effects and student performance (that is, when worse students are assigned to better teachers). We summarize different combinations of grouping and assignment mechanisms that might be encountered in educational settings in Table 1, along with acronyms that we use in the remainder of the paper.

It is important to recognize that a mixture of these grouping and assignment methods can be used in any given district or even within a given school. However, for the sake of clarity in understanding and evaluating the performance of various estimators, we keep the scenarios distinct when we conduct our simulations and assume that all schools simultaneously use the same process.

Generally, the random assignment of groups of students (regardless of how the groups may be formed) to available teachers is not a violation of either strict exogeneity or heterogeneity exogeneity and thus may not cause problems for standard estimation methods. The students may be grouped using dynamic or static assignment provided the teachers are randomly assigned to the groups. Of course, grouping may have other consequences, such as inducing correlation within classrooms in the unobserved factors affecting performance. But this is different from failure of exogeneity.

The systematic assignment of high-performing students to either high- or low-performing teachers, on the other hand, can violate exogeneity assumptions. Dynamic grouping coupled with positive or negative assignment virtually always causes failure of strict exogeneity, because if the teacher assignment is correlated with past scores, then teacher assignment must be correlated with the innovations (errors) that affect past scores. In addition, if student heterogeneity c_i exists then dynamic grouping with nonrandom assignment violates heterogeneity exogeneity, too: part of past performance depends on c_i. It should be noted that Dieterle et al. (unpublished) find

empirical evidence to suggest that ability grouping with positive assignment may occur to a nontrivial degree in school systems.

The two cases of *static* grouping differ in important ways. For example, suppose students are grouped on a baseline score upon entry to school and then assigned to teachers nonrandomly in all subsequent grades. While this is a case of nonrandom assignment, for some estimation approaches there is no violation of relevant exogeneity assumptions. As an illustration, in the gain score equation (9), the baseline score does not appear. Therefore, even if teacher assignment is determined by the base score, if it is independent of the student heterogeneity c_i and the errors e_{it}, then pooled OLS estimation consistently estimates β_0 (and the other parameters). Of course, this assumes that $\lambda = 1$ has been correctly imposed. If $\lambda < 1$, then the gain-score equation effectively omits the lagged test score, and this lagged score will be correlated with the base score, causing bias in any of the usual estimators applied to (9).

Static assignment based on c_i causes problems for estimating equations such as (9) unless $\pi_t c_i$ is removed from the equation. When π_t is constant, the fixed effects and first-differencing transformations do exactly that. Therefore, assigning students to teachers based on the student heterogeneity does not cause problems for these types of estimators applied to (9). But other estimators, particularly POLS and RE, will suffer from omitted variable bias because E_{it} is correlated with c_i. Static assignment based on student growth also causes problems for DOLS because DOLS ignores c_i in estimating (10).

Until now, we have focused on the assignment of students to teachers within schools. Another key consideration, however, is the sorting of students and teachers across schools. If higher achieving students are grouped within certain schools and lower achieving students in others, then the teachers in the high-achieving schools, regardless of their true teaching ability, will have higher probabilities of high-achieving classrooms. Similarly, if higher ability teachers are grouped within certain schools and lower ability teachers in others, then students in the schools with better teachers will realize higher gains. If both high ability teachers and high performing students are then grouped together within schools, the nonrandom sorting issue is exacerbated.

In designing our simulation scenarios, we therefore consider three distinct "school sorting" cases. In Case 1, both students and teachers are randomly placed in schools. Thus there is no systematic difference in average test scores or average true teacher effects across schools.

In Case 2, students are sorted into schools according to their baseline levels of learning but teachers are still randomly placed in schools. Thus there is a significant difference in average test scores across schools but not in average teacher effects. In Case 3, students are randomly placed in schools but teachers are sorted into schools based on their true effects. Thus, there are systematic differences in average teacher effects across schools but not in average test scores.

In our investigation of the performance of various estimators under different sorting, grouping, and assignment scenarios, we focus on how well the estimators meet the needs of policymakers, considering how VAM-based measures of teacher effectiveness might typically be used in educational settings. If districts wish only to rank teachers in order to identify those who are high or low performing, then estimators that come close to getting the rankings right are the most desirable. For the purposes of structuring rewards and sanctions or identifying teachers in need of professional development, districts may wish primarily to distinguish high and low performing teachers from those who are closer to average; if so, it is important that the estimators accurately classify teachers whose performance falls in the tails of the distribution. If, on the other hand, districts wish to know how effective particular teachers are compared with, say, the average, then the teacher effect estimates themselves are of primary importance. Our study investigates the performance of various estimators with respect to all three criteria, using summary measures described in the next section.

5. Methods

Our empirical investigations consist of a series of simulations in which we use generated data to investigate how well each estimator recovers true effects under different scenarios. These scenarios are captured in data generating processes (DGPs) that vary the mechanisms used to assign students to teachers in the ways discussed in the previous section. To data generated from each DGP, we apply the set of estimators discussed in Section 3. We then compare the resulting estimates with the true underlying effects.

5.1. Data Generating Processes

To isolate fundamental problems, we restrict the DGPs to a relatively narrow set of idealized conditions. We assume that test scores are perfect reflections of the sum total of a child's learning (that is, with no measurement error) and that they are on an interval scale that remains constant across grades. We assume that teacher effects are constant over time and that unobserved child-specific heterogeneity has a constant effect in each time period. We assume

there are no time-varying child or family effects, no school effects, no interactions between students and teachers or schools, and no peer effects. We also assume that the GDL assumption holds—namely, that decay in schooling effects is constant over time. In addition, we assume no serial correlation. Finally, there are no time effects embedded in our DGPs.

Our data are constructed to represent three elementary grades that normally undergo standardized testing in a hypothetical district. To mirror the basic structural conditions of an elementary school system for, say, grades 3 through 5 over the course of three years, we create data sets that contain students nested within teachers nested within schools, with students followed longitudinally over time. Our simple baseline DGP is as follows:

$$\begin{aligned} A_{i3} &= \lambda A_{i2} + \beta_{i3} + c_i + e_{i3} \\ A_{i4} &= \lambda A_{i3} + \beta_{i4} + c_i + e_{i4} \\ A_{i5} &= \lambda A_{i4} + \beta_{i5} + c_i + e_{i5} \end{aligned} \qquad (13)$$

where A_{i2} is a baseline score reflecting the subject-specific knowledge of child i entering third grade, A_{i3} is the achievement score of child i at the end of third grade, λ is a time constant decay parameter, β_{it} is the teacher-specific contribution to growth (the true teacher value-added effect), c_i is a time-invariant child-specific effect, and e_{it} is a random deviation for each student. Because we assume independence of e_{it} over time, we are maintaining the common factor restriction in the underlying cumulative effects model. We assume that the time-invariant child-specific heterogeneity c_i is correlated at about 0.5 with the baseline test score A_{i2}.[7]

In the simulations reported in this paper, the random variables A_{i2}, β_{it}, c_i, and e_{it} are drawn from normal distributions, where we adjust the standard deviations to allow different relative contributions to the scores. It is somewhat challenging to anchor our estimates of teacher effect sizes to those in the literature, however, because reported teacher-related variance components range from as low as 3 percent to as high as 27 percent and obtained through different estimation methods (e.g., Nye et al. 2004, McCaffrey et al. 2004, Lockwood et al. 2007). Estimates in the smaller end of the range—i.e., around 5 percent—are more frequently reported. In our own investigations of data from a set of districts, however, we found rough estimates of teacher effects tending toward 20 percent of the total variance in gain scores but highly variable across

[7] Other work by the authors (Reckase et al., unpublished) finds that when test scores are generated as in (13) such correlation—which seems realistic—is necessary to achieve data that conform to the parameter estimates derived from observed achievement distributions.

districts. Thus in our simulations, we explore two parameterization schemes. In the first, the standard deviation of the teacher effect is .25, while that of the student fixed effect is .5, and that of the random noise component is 1, each representing approximately 5, 19, and 76 percent of the total variance in gain scores, respectively, In the second, the standard deviation of the teacher effect is .6, while that of the student fixed effect is .6, and that of the random noise component is 1, representing approximately 21, 21, and 58 percent of the total variance in gain scores, respectively, Thus, in the latter scenario, teacher effects are relatively more important and should be easier to estimate.

Our data structure has the following characteristics that do not vary across simulation scenarios:

- 10 schools
- 3 grades (3^{rd}, 4^{th}, and 5^{th}) of scores and teacher assignments, with a base score in 2^{nd} grade
- 4 teachers per grade (thus 120 teachers overall)
- 20 students per classroom
- 4 cohorts of students
- No crossover of students to other schools

To create different scenarios, we vary certain key features: the sorting of students and teachers into schools, the grouping of students into classes, the assignment of classes of students to teachers within schools, and the amount of decay in prior learning from one period to the next. Within each of the three school-sorting cases outlined in the previous section, we generate data using each of the 10 different mechanisms for the assignment of students outlined in Table 1.[8] Finally, we vary the decay parameter λ as follows: (1) $\lambda = 1$ (no decay or complete persistence) and (2) $\lambda = .5$ (fairly strong decay).[9] Thus, we explore $3 \times 10 \times 2 = 60$ different scenarios in this paper.[10] We use 100 Monte Carlo replications per scenario in evaluating each estimator.

5.2. Methods for Estimating Teacher Effects

[8] We introduce a small amount of noise into each grouping process.
[9] Rough estimates of λ in real data cover a wide range of values. Andrabi et al. (2009) find persistence rates of .5 or lower in Pakistani schools.
[10] Several more scenarios that relax various assumptions are added in sensitivity analyses discussed in a later section of the paper.

We estimate the teacher effects using modified versions of the estimating equations (5) and (9). The modified equations reflect the simplifications determined by our DGPs. Specifically, we remove the time-varying intercept because our data have no time effects, we have no time-varying child and family effects, and we assume that $\pi_t = 1$:

$$\Delta A_{it} = E_{it}\beta_0 + c_i + e_{it} \quad (14)$$

$$A_{it} = \lambda A_{i,t-1} + E_{it}\beta_0 + c_i + e_{it} \quad (15)$$

$$\Delta A_{it} = \lambda \Delta A_{i,t-1} + \Delta E_{it}\beta_0 + \Delta e_{it} \quad (16)$$

where E_{it} is the vector of 119 teacher dummies (with one omitted because every estimation method includes an intercept, either explicitly or by accounting for c_i).

For each of the 100 iterations pertaining to one DGP, we estimate effects for each teacher using one of six estimation methods discussed in Section 3: POLS, RE, and FE applied to (14), POLS applied to (15) (which we have called DOLS), Arellano and Bond (AB) applied to (16), and the average residual (AR) approach, which is based on (15) but only nets out the lagged test score from the current test score. We use the statistical software Stata for all data generation and estimation.

5.3. Summary Statistics for Evaluating the Estimators

For each iteration and for each of the six estimators, we save the estimated individual teacher effects, which are the coefficients on the teacher dummies, and also retain the true teacher effects. To study how well the methods uncover the true teacher effects, we adopt some simple summary measures. The first is a measure of how well the estimates preserve the rankings of the true effects. We compute the Spearman rank correlation, $\hat{\rho}$, between the estimated teacher effects, $\hat{\beta}_j$, and the true effects, β_j, and report the average $\hat{\rho}$ across the 100 iterations.

Second, we compute two measures of misclassification. The first is the percentage of above average teachers (in the true quality distribution) who are misclassified as below average in the distribution of estimated effects. The second focuses on the tails of the quality distribution. We determine which teachers are estimated to be in the bottom 20 percent and then display the proportion of teachers at each percentile of the true effect distribution who are classified in this category using graphs.

6. Simulation Results

6.1. Random Sorting of Students and Teachers across Schools (Case 1) and No Decay, with Small Teacher Effects

We first discuss the findings for the case in which students and teachers are randomly sorted into schools and $\lambda = 1$; these are shown in the left side of Table 2. The underlying parameterization scheme used here is the one in which teacher effects represent only five percent of the total variance in gain scores—a percentage frequently reported in literature. Each cell in Table 2 contains two numbers specific to the particular estimator-scenario combination. The first is the average rank correlation between the estimated and true teacher effects over the 100 replications. The second is the average percentage of above average teachers who are misclassified as being below average.

We expect all estimators to work well when students and teachers are both randomly assigned to classes – the RG-RA scenario defined in Table 1. Of course, the estimated teacher effects still contain sampling error, and so we do not expect to rank or classify teachers perfectly using these estimates. We find that DOLS, AR, POLS, and RE yield rank correlations near .8 or higher, with RE producing a rank correlation of about .89. FE and AB have rank correlations well under .7, with the correlation for AB being the worst at .6. That the FE and AB estimators yield notably lower correlations is not terribly surprising given that they unnecessarily remove a student effect in this scenario; in addition, AB unnecessarily estimates a coefficient on the lagged test score.

The DOLS, AR, POLS, and RE estimators are also better at classifying the teachers than the other two methods. RE incorrectly classifies an above average teacher as being below average about 14% of the time; the misclassification rate for DOLS is somewhat worse at 21%. The misclassification rates for FE and AB, on the other hand, are 26%. Clearly, the estimation error in the teacher effects using FE and AB has important consequences for using those estimates to classify teachers. The potential for misclassification is explored further in Figure 1 for selected scenarios and estimators. The true teacher percentile rank is represented along the x-axis, and the y-axis represents the proportion of times in which a teacher at a particular true percentile is classified in the bottom quintile of the distribution on the basis of his or her estimate. Thus, a perfect estimator would produce the step function traced on the graph, with y = 1 when x ranges from 0 to 20 and y = 0 when x ranges from just above 20 to 100. Part *a* of Figure 1shows the superiority of DOLS, POLS, and RE over FE in the RG-RA scenario with lambda equal to one. However, it should be noted that even for these estimators under these

idealized conditions, identification of the "worst" teachers appears subject to a nontrivial amount of error.

Taken as a whole, these findings indicate that RE is preferred under RG-RA with no decay, something that econometric theory leads us to expect because RE is the (asymptotically) efficient estimation method. However, POLS produces very similar results, AR results are only slightly worse, and DOLS is fairly similar, as well, despite the fact that AR and DOLS are technically inconsistent because of the presence of the unobserved student effect and its correlation with lagged achievement. DOLS does slightly worse than AR here probably because DOLS unnecessarily nets out the lagged test score from the teacher assignment, resulting in more estimation noise.

Nonrandom grouping mechanisms for students have relatively minor consequences for RE, DOLS, AR, and POLS provided the teachers are *randomly assigned to classrooms* – whether the students are grouped according to their prior scores (DG-RA), baseline scores (BG-RA), or heterogeneity (HG-RA).[11] Generally, nonrandom grouping of students causes POLS, and RE to do less well in terms of precision – especially when grouping is based on student heterogeneity – most likely because the student grouping induces cluster correlation within a classroom.[12] Nevertheless, they continue to yield relatively high correlations, ranging from .81 to .86. The methods that remove the student effect, FE and AB, continue to do a much worse job in ranking and also classifying teachers, even under heterogeneity grouping. Their performance is particularly poor for dynamic grouping with random assignment. RE is the best choice in scenarios in which $\lambda = 1$ and assignment is random.

When teachers are *nonrandomly assigned* to classrooms the properties of the estimation procedures change markedly – and it depends critically on the nature of the nonrandom assignment. When dynamic grouping is used and better students are assigned to the better teachers (i.e., the DG-PA scenario), DOLS is the preferred estimation approach for both the DG-PA and DG-NA scenarios because it directly controls for the assignment mechanism. Across all the DG scenarios, DOLS is the most robust estimator. The AR method is clearly inferior to

[11] Note that all random assignment scenarios are shown in shaded cells in the tables.

[12] The consequences of clustering are easy to understand with a method such as POLS. The POLS estimates are simply within-teacher averages of the student gain scores. The sampling method that gives the most precise estimates of a population average is random sampling, where observations are independent (as in the RG-RA design). With cluster correlation each new student effectively adds less information because that student's outcome is correlated with other students' outcomes in the cluster.

DOLS for both DG-PA and DG-NA because it does not partial out the effect of lagged test scores from the teacher assignment dummies. POLS and RE – which both leave c_i in the error term – do well in ranking teachers in the DG-PA scenario—exceeding even the performance of DOLS—but poorly in the DG-NA scenario. The switch from positive to negative assignment has the opposite effect on FE. This is because systematic biases push the estimates in opposite directions.[13] Under DG-NA, the estimators that remove the student-level heterogeneity – FE and AB – perform especially poorly, producing a negative rank correlation between the estimated and true teacher effects ($-.32$ for FE and $-.07$ for AB) and misclassifying large numbers of the above-average teachers as below average (56% for FE and 46% for AB). The poor performance of FE is highlighted in Figure 1, Part *b*, which vividly illustrates how the best teachers are more likely to be classified as underperforming than the worst ones. In this type of scenario—with students grouped on the basis of past test scores and assigned to teachers whose performance tends to match their own—these procedures will not be helpful in distinguishing among teachers.

Nonrandom teacher assignment coupled with either of the two *static grouping* mechanisms also poses challenges. When $\lambda = 1$ and static grouping of students is based on the baseline score, the DOLS and AR estimators fluctuate least across the two scenarios with nonrandom assignment, although DOLS performs slightly better. AR is systematically biased because it does not net out the base score, A_{i2}, from the teacher assignment. DOLS is systematically biased because it effectively controls for the wrong explanatory variable, $A_{i,t-1}$, when it should control for the base score, A_{i2}. This problem with DOLS can be seen, with $\lambda = 1$, by writing A_{it} as a function of all past inputs, shocks, and the initial value. The resulting equation includes A_{i2} with a time-varying coefficient. We can think of $A_{i,t-1}$ acting as an imperfect proxy forA_{i2}. POLS and RE fluctuate greatly across the two scenarios, indicating a systematic bias. The bias is introduced through correlation between the base score and the student fixed effect—if these are uncorrelated,[14] POLS and RE would be consistent and better performing. FE and AB remain fairly stable across the BG-PA and BG-NA scenarios but again

[13] Further analysis reveals that the estimates may look good when the distance from the mean teacher is inflated and poor when that distance is compressed. These biases create the large swing from good to poor when we go from positive to negative assignment. The effect of the bias is highly dependent on the context and therefore cannot be predicted when schools are using a mixture of positive and negative assignment.

[14] This is admittedly an unlikely possibility, but it was one that we explored in analyses not shown. Although some correlations and misclassification rates were affected, the general patterns revealed in the findings were the same whether the base score and student fixed effect were correlated or uncorrelated.

perform poorly. The second type of static grouping mechanism (HG) combines students based on the value of c_i - the time invariant student-specific growth potential. When $\lambda = 1$, c_i is a permanent component of the gain score. That is, c_i is added, in each period, to the previous score. When the students with the highest growth potential are grouped with the best teachers (HG-PA), the bias in DOLS, AR, POLS, and RE (estimators that ignore c_i) leads them to rank and classify the teachers well. But negative assignment causes them to do much worse. In fact, in the HG-NA scenario, no estimator does very well – the highest rank correlation is .62 (FE) and the lowest misclassification rate is 26% (FE). Figure 1.c illustrates the decline in performance of RE and DOLS relative to the scenario depicted in Part *a*. Theoretically, the FE estimator is the most efficient estimator among those that place no restrictions on the relationship between c_i and the teacher dummies E_{it}. But the consistency of FE (and AB) is of small comfort, as it does not outperform the estimators that effectively treat c_i and the teacher dummies E_{it} as being uncorrelated along the dimensions that matter most: ranking and classifying. So far, even though we have discussed only the case of nonrandom sorting of students and teachers across schools and no decay, we can summarize some useful insights. First of all, the findings show that even under these idealized conditions, certain estimators perform very poorly under certain assignment mechanisms – even some estimators that effectively use the fact that $\lambda = 1$ in estimation. Estimators that are intended to be robust to static assignment do poorly under dynamic assignment.

A useful finding, in looking across all assignment mechanisms is that DOLS does best: it is superior under dynamic grouping and still has value for ranking teachers under static grouping. We can understand the relatively good performance of DOLS under the various dynamic grouping scenarios by noting that if the DGP did not include a student effect, the DOLS estimator would be consistent across all assignment mechanisms associated with this scenario. Namely, the teacher dummies E_{it} are correlated with $A_{i,t-1}$ but the latter is controlled for in the DOLS regression. AR includes the lagged test score in the first-step regression but does not partial out its correlation with the teacher dummies. Because AB uses the differenced equation and instruments for the lagged gain score using test scores dated two or more periods ago, it also does not properly partial out the lagged level of the test score from teacher assignment. AB would likely work better if assignment in grade *t* depended only on the test score two grades earlier, but this assignment scenario seems unrealistic. POLS, RE, and FE do not control for the

lagged achievement score because they use the gain score as the dependent variable and omit the lag. POLS and RE – which both leave c_i in the error term – suffer from an omitted variable problem because assignment is based on the lagged test score and the lagged score is positively correlated with c_i. In the DG-PA case, the resulting bias in estimating the teacher effects by POLS or RE actually helps with ranking the students, but it hurts in the DG-NA case. DOLS, on the other hand, exhibits more or less the same behavior whether assignment is RA, PA, or NA.

6.2. Random Sorting of Students and Teachers across Schools (Case 1) and Strong Decay with Small Teacher Effects

The performance of several estimators deteriorates substantially when we change the value of λ from 1 to .5. The right side of Table 2 shows simulation findings when students and teachers are randomly assigned to schools and $\lambda = .5$. Importantly, because POLS, RE, and FE act as if $\lambda = 1$, these estimators are now applied to an equation with misspecified dynamics, regardless of the assignment mechanism. Because POLS, RE, and FE use the gain score as the dependent variable, an omitted variable, $A_{i,t-1}$ in equation (14) will have a coefficient of $-.5$ on it; this is important to remember in interpreting the findings.

Dynamic misspecification has a large effect on the precision of the estimates. Compared with the $\lambda = 1$ DGP, the rank correlations for POLS, RE, and FE are substantially worse when $\lambda = .5$. For example, even in the RG-RA scenario, the rank correlation for RE is only .55, down from .89. The misclassification rate is 35% compared with 14% when $\lambda = 1$. The impact on POLS and RE for misclassifying low-performing teachers is seen in Figure 1.d[15].

When coupled with dynamic assignment, dynamic misspecification has very serious consequences for all estimators with the notable exception of DOLS. AR is also unaffected but remains inferior to DOLS. When the best students are matched with the best teachers (DG-PA), POLS, RE, FE, and AB actually produce rank correlations that are either close to zero or negative and have 52% or greater misclassification rates. The striking effects for misclassification at the tails of the quality distribution are visible in Figures 1.e.

Dynamic misspecification also has consequences in the case of static assignment. POLS and RE do a very poor job of ranking and classifying students in the BG-PA case. The rank correlation is only .2. FE and AB do better in this case, but both are clearly inferior to DOLS.

[15] Because $\lambda < 1$, the composite errors in the random effects estimation have negative serial correlation, causing the estimated variance of c_i to be negative. In such cases, Stata sets the variance to zero, which leads to the POLS estimate. This happens across all simulations, so POLS and RE are identical.

With negative assignment, POLS and RE do substantially better but are always inferior to DOLS. Even when grouping is based on heterogeneity, DOLS generally does better than other estimators. DOLS does not do particularly well in the HG-NA setting but at least it produces a nontrivial rank correlation (.5). As we saw in the case when $\lambda = 1$, no estimator works very well in the HG-NA case, and all have extremely high misclassification rates, most of which display more error than chance.

Taken as a whole, the simulations for Case 1—i.e., when both students and teachers are randomly assigned to schools, combined with small teacher effects, point to several conclusions. While DOLS is not uniformly better across all of the grouping, assignment, and decay assumptions, it is nearly so. DOLS is easily preferred under dynamic grouping: looking across the different assignment mechanisms and both values of λ, no estimator is even a close second. The performance of DOLS is stable across values of λ. The other estimators show much more sensitivity to the value of λ. The robustness of DOLS makes it the recommended approach among the group of estimators considered in this study. However, we should note that the potential for misclassification in these simple DGPs, even using DOLS, can approach levels that might be considered unacceptable for policy purposes.

6.3. Large Teacher Effects

We also conducted a set of simulations where teacher effects represent a much larger relative share of the total variance in student gains. These are reported in Table 3, as well as Figure 2. As to be expected, when the size of the teacher effects is raised relative to the student effect and shock, rank correlations improve and misclassification rates decline somewhat. The same overall patterns observed in the "small" teacher effects case continue to hold. The relative superiority of DOLS over AR in the DG scenarios and over POLS and RE in the scenarios with strong decay is still evident when teacher effects are large. The FE and AB estimators improve their rank correlations in many scenarios when teacher effects are large but remain the least effective estimators overall. Although concerns over inaccuracy in the estimates and rankings are mitigated when teacher effects are large, the same lessons regarding which estimator to use in particular contexts apply, and the overall conclusion that DOLS is more robust across scenarios holds.

6.4. Sensitivity Analyses

We subjected our simulations to several sensitivity analyses. First, we looked at the impact of nonrandom sorting of students and teachers across schools. These different sorting scenarios did little to affect the general patterns described above, indicating that the primary threat to the estimation of teacher effects stems from within-school assignment to teachers. However, some changes in the correlations and misclassification rates were evident for most estimators in most scenarios. POLS and RE deteriorated slightly when students were nonrandomly sorted across schools. FE and AB deteriorated substantially when teachers were nonrandomly sorted across schools.

We also ran a full set of simulations with $\lambda = .75$ (more moderate decay), without any surprises. This implies a less severe form of dynamic misspecification for estimators such as POLS and RE than the $\lambda = .5$ case. It is not surprising that the performance of POLS and RE was essentially between the $\lambda = 1$ and $\lambda = .5$ cases. The DOLS estimator was hardly affected by the value of λ. Nor was AR, but it was still generally outperformed by DOLS.

We also added classical measurement error to the test scores in our DGPs. In addition, we ran simulations in which serial correlation was introduced in the errors (i.e., relaxing the common factor restriction). While these complications did not affect the overall patterns described, they generally served to depress rank correlations slightly and to increase misclassification rates for POLS, RE, FE, and AB (they were particularly damaging to AB when there was no decay). For DOLS and AR, the effects of measurement error were mixed. Serial correlation actually seemed to improve DOLS and AR slightly[16]. Given that serial correlation can be picked up by including a lagged test score, this is not terribly surprising. Remember, we are not out to estimate the coefficient on the lagged test score but only to rank teachers based on the estimated VAMs.

Finally, we examined the performance of the estimators when student mobility across schools was present. When we allowed 10 percent of students to switch schools in each year, FE and AB improved overall but still suffered in the dynamic grouping-nonrandom assignment scenarios and were still not as robust as DOLS. For all sensitivity analyses, details are available from the authors upon request.

[16] Tables containing the additional simulation findings are available on request from the authors.

7. Conclusions and Future Directions

Simulated data with known properties permits the systematic exploration of the ability of various estimation methods to recover the true parameters used to generate the data—in our case teacher effects. This study has taken the first step in evaluating different value-added estimation strategies under conditions in which they are most likely to succeed. Creating somewhat realistic but idealized conditions facilitates the investigation of issues associated with the use of particular estimators. If they perform poorly under these idealized conditions, they will almost certainly do worse in real settings.

Our main finding is that no one method is guaranteed to accurately capture true teacher effects in all contexts even under these relatively idealized conditions, although some are more robust than others. Because we consider a variety of DGPs, student grouping mechanisms, and teacher assignment mechanisms, it is not surprising that no single method works well in all hypothetical contexts. Both the teacher assignment mechanism and the nature of the dynamic relationship between current and past achievement play important roles in determining how well the estimators function.

A dynamic specification estimated by OLS—what we have called DOLS—was, by far, the most robust estimator across scenarios. Only in one scenario—heterogeneity-based grouping with negative assignment—did it fail to produce useful information with regard to teacher effects. However, none of our estimators was able to surmount the problems posed by this scenario—not even estimators designed to eliminate bias stemming from unobserved heterogeneity—and it is perhaps a less realistic scenario than others we considered.

In all other situations, DOLS provided estimates of some value. The main strength of this estimator lies in the fact that, by including prior achievement on the right-hand side, it controls either directly or indirectly for grouping and assignment mechanisms. In the case of dynamic grouping coupled with non-random assignment, it explicitly controls for the potential source of bias. In the case of baseline and heterogeneity grouping, the effect of controlling for prior achievement is less direct but still somewhat effective in that both those grouping mechanisms are correlated with prior achievement.

These findings suggest that choosing estimators on the basis of structural modeling considerations may produce inferior results by drawing attention to relatively unimportant concerns and away from key concerns. The DOLS estimator is never the prescribed approach

under the structural cumulative effects model with a geometric distributed lag (unless there is no student heterogeneity), yet it is often the best estimator. One can think of the DOLS estimator as a regression-based version of a dynamic treatment effects estimator. That is not to say that the general cumulative effects model is incorrect. It merely reflects the fact that efforts to derive consistent estimators by focusing on particular concerns of structural modeling (e.g., heterogeneity, endogenous lags) may obscure the fact that controlling for the assignment mechanism even in specifications that contain other sources of endogeneity is essential. Approaches that attend to less important features of the structural model, when coupled with nonrandom assignment, may yield estimators that are unduly constrained and thus poorly behaved. The poor performance of the AB estimator exemplifies this. By differencing for heterogeneity and using instrumental variables to remove bias from the estimation of lambda, it loses much of its ability to estimate teacher effects precisely. The findings in this paper suggest that flexible approaches based on dynamic treatment effects (for example, Lechner, 2008;, Wooldridge, 2010, Chapter 21) may be more fruitful than those based on structural modeling considerations.

Finally, despite the relatively robust performance of DOLS, we find that even in the best scenarios and under the simplistic and idealized conditions imposed by our data generating process, the potential for misclassifying above average teachers as below average or for misidentifying the "worst" or "best" teachers remains substantial, particularly if teacher effects are relatively small. Applying the commonly used estimators to our simplified DGPs results in misclassification rates that range from at least five to more than 50 percent, depending upon the estimator and scenario.

It is clear from this study that certain VAMs hold promise: they may be capable of overcoming many obstacles presented by non-random assignment and yield valuable information, providing assignment mechanisms are known or can be deduced from the data. Our findings indicate that teacher rankings can correlate relatively well with true rankings in certain scenarios and that, in some cases, misclassification rates may be relatively low. Given the context-dependency of the estimators' ability to produce accurate results, however, and our current lack of knowledge regarding prevailing assignment practices, VAM-based measures of teacher performance, as currently applied in practice and research, must be subjected to close scrutiny regarding the methods used and interpreted with a high degree of caution.

Methods of constructing estimates of teacher effects that we can trust for high-stakes evaluative purposes must be further studied, and there is much left to investigate. This paper does not address the degree to which test measurement error, school effects, time-varying teacher effects, different types of interactions among teachers and students, and compensating or reinforcing contemporaneous family effects alter the performance of the estimators. Finally, diagnostics are needed to identify the structure of decay and prevailing teacher assignment mechanisms. If contextual norms with regard to grouping and assignment mechanisms can be deduced from available data, then it may be possible to determine which estimators should be applied in a given context. Other work by the authors and others (e.g., Dieterle et al. unpublished) finds that dynamic grouping is present in a nontrivial number of schools and also finds suggestive evidence of positive assignment, offering further evidence of the usefulness of DOLS.

Clearly, although value-added measures of teacher performance hold some promise, more research is needed before they can confidently be implemented in high-stakes policies. Our findings suggest that teacher effect estimates constructed using DOLS may be useful in answering research questions that employ them in regression specifications. The degree of error in these estimates, however, makes them *less trustworthy* for the specific purpose of evaluating individual teachers. It may be argued that including these measures in a comprehensive teacher evaluation along with other indicators could provide beneficial information and represent an improvement over the status quo. However, it would be unwise to use these measures as the sole basis for sanctions. Even if such measures are released to the public simply as information—as has recently been the case in Los Angeles and New York City—the potential for inaccuracy, and thus for damage to teachers' status and morale, creates risks that could outweigh the benefits. If such measures are accurate, then publicizing or attaching incentives to them may motivate existing teachers to increase efforts or induce individuals with high performance potential into the teaching profession. If, however, such measures cannot be trusted to produce fair evaluations, existing teachers may become demoralized and high potential individuals considering teaching as a profession may steer away from entering the public school system.

Given that the accuracy of VAM-based measures of teacher performance can vary considerably across contexts and that the potential for bias if particular methods are applied to

the wrong situations is nontrivial, we conclude that it is premature to attach high stakes to these measures until their properties have been better understood.

References

Aaronson, D., Barrow, L., Sander, W. (2007) Teachers and Student Achievement in the Chicago Public High Schools. *Journal of Labor Economics*, 25(1), 95-135.

Andrabi, T., Das, J., Khwaja, A., and Zajonc, T. (2009), Do Value-Added Estimates Add Value? Accounting for Learning Dynamics, HKS Faculty Research Working Paper Series RWP09-034, John F. Kennedy School of Government, Harvard University, http://dash.harvard.edu/handle/1/4435671, accessed on 5/15/12.

Arellano, M. & Bond, S. (1991) Some Tests of Specification of Panel Data: Monte Carlo Evidence and an Application to Employment Equations. *The Review of Economic Studies*, 58, pp.277-298.

Ballou, D., Sanders, W., & Wright, P (2004), "Controlling for Student Background in Value-Added Assessment of Teachers," *Journal of Educational and Behavioral Statistics* 29, 37-65

Blundell, R. & Bond, S. (1998) Initial Conditions and Moment Restrictions in Dynamic Panel Data Models. *Journal of Econometrics*, 87, 11-143.

Boardman, A. & Murnane, R. (1979) Using Panel Data to Improve Estimates of the Determinants of Educational Achievement, *Sociology of Education*, 52, 113-121.

Buddin, R. (2011) Measuring Teacher and School Effectiveness at Improving Student Achievement in Los Angeles Elementary Schools: http://mpra.ub.uni-muenchen.de/31963/MPRA Paper No. 31963, accessed on 5/15/12.

Chetty, R., Freidman, J., & Rockoff, J. (2011) "The Long-Term Impacts of Teachers: Teacher Value-Added and Student Outcomes in Adulthood." NBER Working Paper 17699.

Dieterle, S., Guarino, C., Reckase, M., & Wooldridge, J. (unpublished draft) *How do Principals Group and Assign Students to Teachers? Finding Evidence in Administrative Data and the Implications for Value-added.*

Downey, D., Hippel, P., & Broh, B. (2004) Are Schools the Great Equalizer? Cognitive Inequality during the Summer Months and the School Year, *American Sociological Review*, 69(5), 613-635.

Entwistle, D. & Alexander, K. (1992) Summer Setback: Race, Poverty, School Composition, and Mathematics Achievement in the First Two Years of School, *American Sociological Review*
Vol. 57, No. 1 (Feb., 1992), pp. 72-84

Hanushek, E. "The Economics of Schooling: Production and Efficiency in the Public Schools," *Journal of Economic Literature*, XXIV (3): 1141-78, 1986.

Hanushek, E. "Conceptual and Empirical Issues in the Estimation of Educational Production Functions," *Journal of Human Resources*, 14(3): 351-388, 1979.

Harris, D., Sass, T., & Semykina, A. (2011) Value-Added Models and the Measurement of Teacher Productivity, Unpublished, Tallahassee, FL: Florida State University.

Kane, T. & Staiger, D. (2008) Estimating Teacher Impacts on Student Achievement: An Experimental Evaluation, Working Paper 14607, National Bureau of Economic Research.

Koedel, C. & Betts, J. (2011) Does Student Sorting Invalidate Value-Added Models of Teacher Effectiveness? An Extended Analysis of the Rothstein Critique. Education Finance and Policy, 6(1), 18-42.

Lechner, M. (2008), Matching Estimation of Dynamic Treatment Models: Some Practical Issues, in *Advances in Econometrics*, Volume 21 (Modeling and Evaluating Treatment Effects in Econometrics). Daniel Millimet, Jeffrey Smith, and Edward Vytlacil (eds.), 289-333-117. Amsterdam: Elsevier, 2008

McCaffrey, D., Lockwood, J.R., Louis, T., & Hamilton, L. (2004) Models for Value-Added Models of Teacher Effects. *Journal of Educational and Behavioral Statistics*, 29(1), pp. 67-101.

Morris, C.N. (1983), "Parametric Empirical Bayes Inference: Theory and Applications," *Journal of the American Statistical Association*, 78, 47-55.

Raudenbush, S. (2009) Adaptive Centering with Random Effects: An Alternative to the Fixed Effects Model for Studying Time-Varying Treatments in School Settings. *Education Finance and Policy*, 4(4), 468-491.

Reckase, M., Ham, E., Guarino, C., & Wooldridge, J. (unpublished draft) *What Can Be Learned from Simulation Studies of Value-Added Models?*

Rothstein, J. (2008) Teacher Quality in Educational Production: Tracking, Decay, and Student Achievement. *NBER Working Paper Series*, Working Paper 14442, http://www.nber.org/papers/w14442.

Sanders, W. & Horn, S. (1994) The Tennessee Value-Added Assessment System (TVAAS): Mixed-Model Methodology in Educational Assessment. *Journal of Personnel in Education*, 8, 299-311.

Todd, P. & Wolpin, K. (2003). On the Specification and Estimation of the Production Function for Cognitive Achievement. *Economic Journal*, 113(485), 3-33.

Santos, F. & Gebeloff, R., Teacher Quality Widely Diffused, Ratings Indicate, http://www.nytimes.com/2012/02/25/education/teacher-quality-widely-diffused-nyc-ratings-indicate.html, accessed on 5/4/12.

US Department of Education (2009) Race to the Top Program: Executive Summary, http://www2.ed.gov/programs/racetothetop/executive-summary.pdf, accessed on 9/8/10.

Value-Added Research Center (2010) NYC Teacher Data Initiative: Technical Report on the NYC Value-Added Model. http://schools.nyc.gov/NR/rdonlyres/A62750A4-B5F5-43C7-B9A3-F2B55CDF8949/87046/TDINYCTechnicalReportFinal072010.pdf, accessed on 5/15/12.

West, M. & Chingos, M. (2010) "Teacher Effectiveness, Mobility, and Attrition in Florida," in Performance Incentives: Their Growing Impact on American K-12 Education, Matthew G. Springer, ed. (Brookings Institution Press), 251-271.

Wooldridge, J.M. (2010), *Econometric Analysis of Cross Section and Panel Data*, 2e. MIT Press: Cambridge, MA.

Zeger, S., Liang, K., & Albert, P. (1988) Models for Longitudinal Data: A Generalized Estimating Equation Approach. *Biometrics*. 44(4), 1049-1060.

Table 1: Grouping and Assignment Acronyms

Acronym	**Process for grouping students in classrooms**	**Process for assigning students to teachers**
RG-RA	Random	Random
DG-RA	Dynamic based on prior test scores	Random
DG-PA	Dynamic based on prior test scores	Positive correlation between teacher effects and prior student scores (better teachers with better students)
DG-NA	Dynamic based on prior test scores	Negative correlation between teacher effects and prior student scores
BG-RA	Static based on baseline test scores	Random
BG-PA	Static based on baseline test scores	Positive correlation between teacher effects and baseline student scores
BG-NA	Static based on baseline test scores	Negative correlation between teacher effects and baseline student scores
HG-RA	Static based on heterogeneity	Random
HG-PA	Static based on heterogeneity	Positive correlation between teacher effects and student fixed effects
HG-NA	Static based on heterogeneity	Negative correlation between teacher effects and student fixed effects

Table 2: Results from 100 replications of Case 1-random sorting of students and teachers across schools. Small teacher effects. Row 1: Average rank correlation. Row 2: Percentage of above average teachers misclassified as below average

Small Teacher Effects	λ=1						λ=.5					
Estimator	DOLS	AR	POLS	RE	FE	AB	DOLS	AR	POLS	RE	FE	AB
Assignment Scenario												
	0.78	0.81	0.88	0.89	0.63	0.60	0.78	0.82	0.55	0.55	0.48	0.60
RG-RA	21%	19%	15%	14%	26%	26%	21%	19%	35%	35%	32%	26%
	0.78	0.81	0.81	0.86	0.59	0.52	0.78	0.81	0.52	0.52	0.42	0.54
DG-RA	21%	19%	19%	16%	27%	29%	21%	19%	34%	34%	32%	29%
	0.77	0.68	0.90	0.90	-0.32	-0.07	0.79	0.70	0.05	0.05	-0.41	-0.09
DG-PA	23%	28%	13%	13%	56%	46%	21%	27%	52%	52%	54%	53%
	0.76	0.62	0.32	0.32	0.74	-0.19	0.78	0.68	0.74	0.74	0.69	0.74
DG-NA	20%	29%	41%	41%	20%	54%	21%	26%	22%	22%	21%	21%
	0.78	0.81	0.84	0.86	0.63	0.60	0.78	0.81	0.54	0.54	0.48	0.61
BG-RA	21%	19%	18%	16%	25%	26%	21%	19%	35%	35%	32%	26%
	0.82	0.78	0.91	0.92	0.63	0.60	0.84	0.81	0.36	0.36	0.50	0.60
BG-PA	19%	22%	12%	12%	24%	26%	17%	19%	43%	43%	30%	26%
	0.73	0.71	0.55	0.65	0.61	0.58	0.69	0.67	0.62	0.62	0.42	0.59
BG-NA	23%	24%	32%	28%	25%	27%	27%	27%	35%	35%	35%	27%
	0.77	0.79	0.82	0.85	0.63	0.59	0.77	0.79	0.55	0.55	0.48	0.60
HG-RA	22%	21%	18%	17%	25%	27%	22%	21%	35%	35%	32%	26%
	0.87	0.86	0.91	0.91	0.63	0.61	0.87	0.87	0.61	0.61	0.45	0.61
HG-PA	16%	16%	12%	12%	25%	26%	15%	16%	33%	33%	34%	26%
	0.51	0.50	0.39	0.55	0.62	0.58	0.50	0.49	0.45	0.45	0.49	0.59
HG-NA	33%	34%	39%	33%	26%	26%	34%	35%	39%	39%	30%	27%

Table 3: Results from 100 replications of Case 1-random sorting of students and teachers across schools. Large teacher effects. Row 1: Average rank correlation. Row 2: Percentage of above average teachers misclassified as below average

Large Teacher Effects	λ=1						λ=.5					
Estimator / **Assignment Mechanism**	**DOLS**	**AR**	**POLS**	**RE**	**FE**	**AB**	**DOLS**	**AR**	**POLS**	**RE**	**FE**	**AB**
	0.94	0.95	0.97	0.97	0.70	0.68	0.94	0.95	0.83	0.83	0.61	0.69
RG-RA	11%	9%	7%	7%	23%	23%	11%	9%	18%	18%	27%	23%
	0.93	0.94	0.94	0.97	0.69	0.66	0.93	0.95	0.81	0.81	0.59	0.68
DG-RA	11%	10%	10%	8%	23%	24%	11%	10%	20%	20%	28%	23%
	0.93	0.78	0.96	0.96	0.32	-0.21	0.93	0.80	0.57	0.57	-0.02	0.50
DG-PA	12%	23%	8%	8%	37%	50%	11%	22%	33%	33%	44%	32%
	0.92	0.84	0.82	0.83	0.76	0.16	0.93	0.88	0.88	0.88	0.72	0.77
DG-NA	11%	18%	19%	19%	20%	40%	11%	16%	15%	15%	21%	19%
	0.94	0.95	0.96	0.97	0.70	0.69	0.94	0.95	0.83	0.83	0.61	0.69
BG-RA	11%	10%	9%	7%	23%	23%	11%	10%	19%	19%	27%	23%
	0.94	0.90	0.97	0.97	0.70	0.69	0.95	0.91	0.75	0.75	0.61	0.69
BG-PA	10%	14%	7%	7%	22%	23%	9%	13%	24%	24%	27%	23%
	0.91	0.93	0.90	0.94	0.69	0.68	0.90	0.92	0.83	0.83	0.56	0.68
BG-NA	13%	12%	14%	11%	23%	23%	14%	13%	21%	21%	30%	24%
	0.93	0.93	0.94	0.96	0.70	0.69	0.93	0.93	0.83	0.83	0.61	0.69
HG-RA	12%	11%	10%	8%	23%	24%	12%	11%	19%	19%	27%	23%
	0.95	0.94	0.96	0.97	0.70	0.69	0.94	0.95	0.83	0.83	0.61	0.69
HG-PA	10%	11%	8%	7%	22%	23%	11%	9%	18%	18%	27%	23%
	0.85	0.85	0.82	0.89	0.69	0.68	0.93	0.95	0.81	0.81	0.59	0.68
HG-NA	17%	18%	20%	15%	22%	23%	11%	10%	20%	20%	28%	23%

Figure 1. Small teacher effect (thick solid=perfect classification, solid=DOLS, dash=POLS, cross=RE, dot=FE, asterisk=AR)

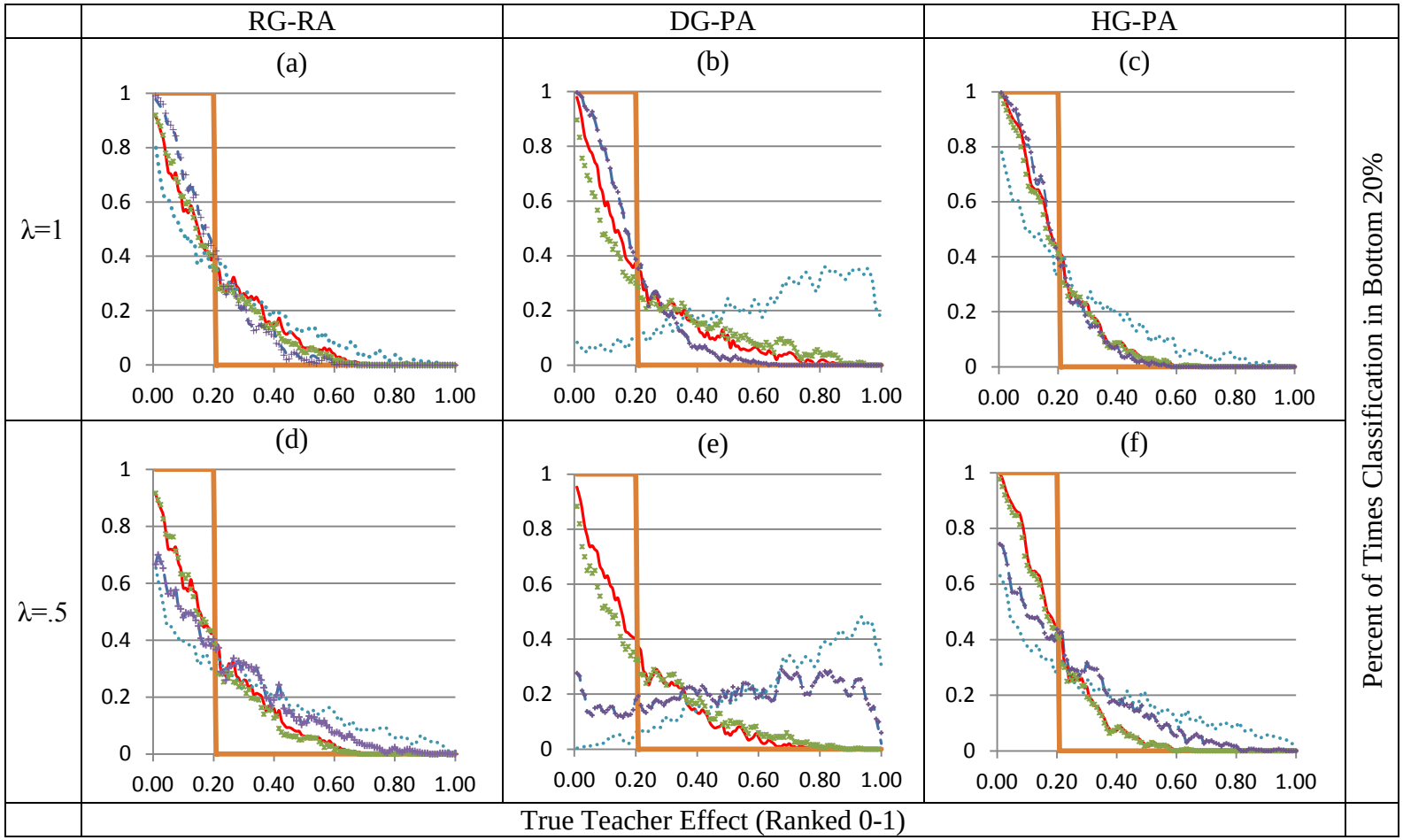

Figure 2. Large teacher effect (thick solid=perfect classification, solid=DOLS, dash=POLS, cross=RE, dot=FE, asterisk=AR)